

A Guideline of Selecting and Reporting **Intraclass Correlation Coefficients** for Reliability Research





Cracking the Code: Providing Insight Into the Fundamentals
of Research and Evidence-Based Practice

2

A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research



Terry K. Koo, PhD^{a,*}, Mae Y. Li, BPS^b

^a Director & Associate Professor, Foot Levelers Biomechanics Research Laboratory, New York Chiropractic College, Seneca Falls, NY

^b DC Candidate, Foot Levelers Biomechanics Research Laboratory, New York Chiropractic College, Seneca Falls, NY

Received 30 July 2015; received in revised form 3 November 2015; accepted 9 November 2015

Key indexing terms:

Reliability and validity;
Research;
Statistics

Abstract

Objective: Intraclass correlation coefficient (ICC) is a widely used reliability index in test-retest, intrarater, and interrater reliability analyses. This article introduces the basic concept of ICC in the content of reliability analysis.

Discussion for Researchers: There are 10 forms of ICCs. Because each form involves distinct assumptions in their calculation and will lead to different interpretations, researchers should explicitly specify the ICC form they used in their calculation. A thorough review of the research design is needed in selecting the appropriate form of ICC to evaluate reliability. The best practice of reporting ICC should include software information, “model,” “type,” and “definition” selections.

Discussion for Readers: When coming across an article that includes ICC, readers should first check whether information about the ICC form has been reported and if an appropriate ICC form was used. Based on the 95% confident interval of the ICC estimate, values less than 0.5, between 0.5 and 0.75, between 0.75 and 0.9, and greater than 0.90 are indicative of poor, moderate, good, and excellent reliability, respectively.

Conclusion: This article provides a practical guideline for clinical researchers to choose the correct form of ICC and suggests the best practice of reporting ICC parameters in scientific publications. This article also gives readers an appreciation for what to look for when coming across ICC while reading an article.

© 2016 National University of Health Sciences.

Reliability

3

- Reliability is defined as the extent to which measurements can be replicated.
- In other words, it reflects not only degree of correlation but also agreement between measurements.
- Mathematically, reliability represents a ratio of true variance over true variance plus error variance.
- As indicated in the calculation, reliability value ranges between 0 and 1, with values closer to 1 representing stronger reliability.
- Historically, Pearson correlation coefficient, paired t test, and Bland-Altman plot have been used to evaluate reliability.

Reliability

4

- However, paired t test and Bland-Altman plot are methods for analyzing agreement.
- Pearson correlation coefficient is only a measure of correlation, and hence, they are non ideal measures of reliability.
- A more desirable measure of reliability should reflect both degree of correlation and agreement between measurements of **continuous data**.
- **Intraclass correlation coefficient (ICC) is such as an index.**

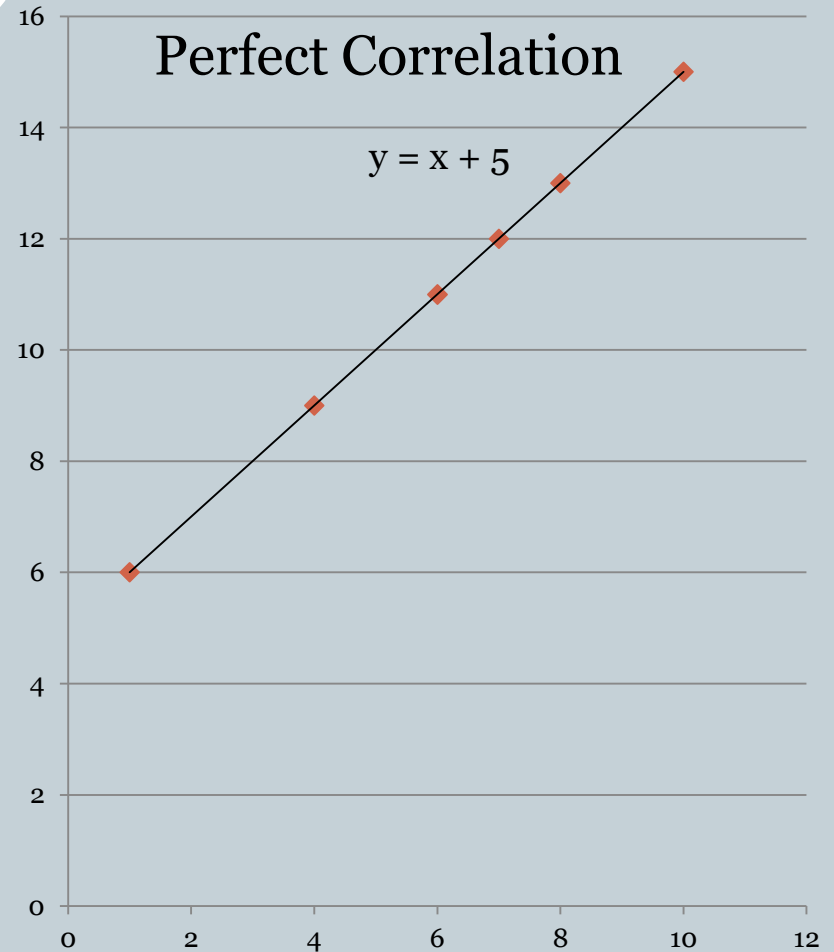
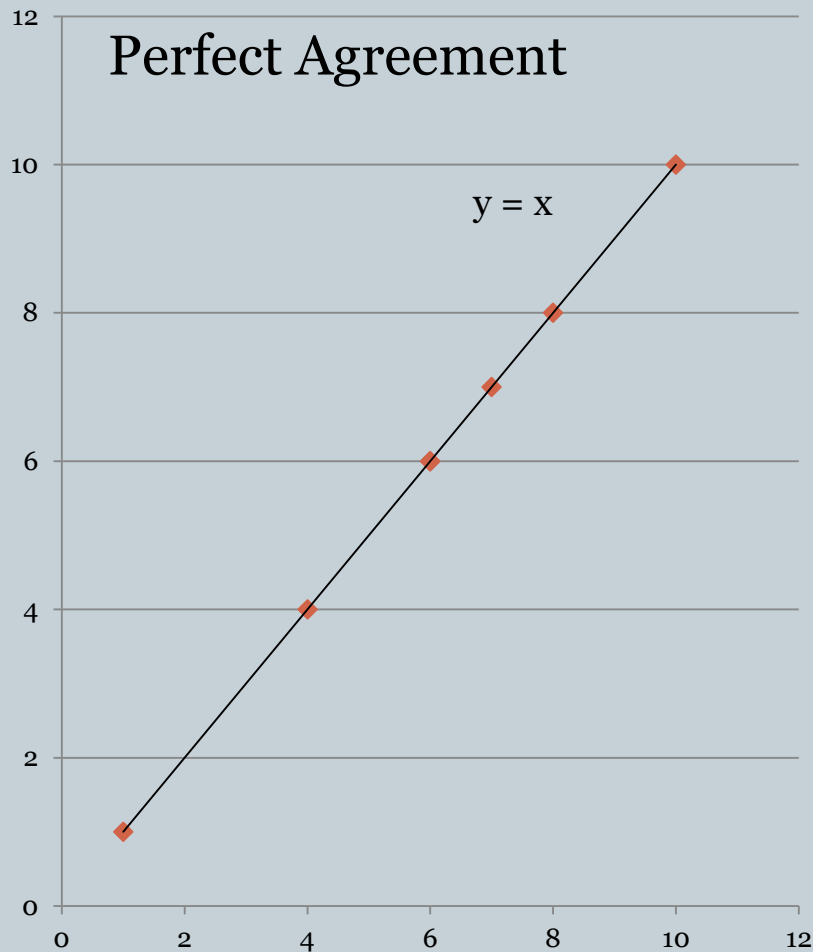
Stop using Pearson r in Reliability in test-retest

5

- The Pearson r share neither their metric nor variance because Pearson's r simply does not measure agreement because it is completely insensitive to changes in scale.
- Pearson r measures the strength of a linear relation between two variables, not the agreement between them.
- We will have perfect agreement only if the points lie along the line of equality, but we will have perfect correlation if the points lie along any straight line.

Agreement VS Correlation

6



Why Confused ICC with Pearson r

7

- The Pearson r is readily available and much easier to understand.
- A second, but much more subtle reason is that some research scientists have applied the Pearson r and an appropriate model of the ICC to the same data set and have obtained very similar results. When there is, in fact, a high level of agreement between any given pair of raters, the r Pearson and an appropriate model of the ICC will indeed produce similar results.
- The point here is that, the r Pearson simply measures the extent to which pairs of raters' scores vary in the same order, *not the extent to which the raters' individual scores actually disagree with each other.*

Definitions of Different Types of Reliability based on the study design

8

- **Interrater reliability**, It reflects the variation between 2 or more raters who measure the same group of subjects.
- **Test-retest reliability**, It reflects the variation in measurements taken by an instrument on the same subject under the same conditions. It is generally indicative of reliability in situations when raters are not involved or rater effect is neglectable, such as self-report survey instrument.
- **Intrarater reliability**, It reflects the variation of data measured by 1 rater across 2 or more trials.
- Issues regarding internal consistency are not addressed here.

Literature Review

- Intra class correlation coefficient was first introduced by **Fisher** in **1954** as a modification of Pearson correlation coefficient.
- **Shrout and Fleiss 1979**, defined six different formulas for calculating the ICC which depend on the purpose of the study, the design of the study and type of measurements taken. The first number designates the “**model**” (1-way/ 2-way), and the second number designates the “**Form**” (single rater / average raters measurements),
- **McGraw and Wong 1996**, defined 10 forms of ICC based on the “**Model**”, the “**Form**” and the “**Definition**” of relationship considered to be important (consistency / absolute agreement).
- **SPSS** provides easy to use tools to measure the ICCs, but the ICCs employed by SPSS is based on McGraw and Wong (1996)

“Models” of the ICC (Shrout and Fleiss)

10

- Model 1 – each subject is assessed by a different set of randomly selected raters. This is rare in reliability studies. **1-way random effects**
- Model 2 – each subject is assessed by each rater, and raters have been randomly selected. **2-way random effects**
- Model 3 – each subject is assessed by each rater, but the raters are the only raters of interest. **2-way mixed effects**

“Forms” of the ICC (Shrout and Fleiss)

11

- The form reflects whether the reliability is to be calculated on a single measurement or by taking the average of 2 or more measurements taken by different raters. In most cases, the form will be 1, however if you want to test whether taking an average of 2 raters' scores improves reliability, you might use form 2,3,4,etc.
- Single measurement = 1
- Average of 2 measurements = 2
- Average of 3 measurements =3

ICC type Description (Shrout and Fleiss)

12

- ICC(1,1) Each subject is assessed by a different set of randomly selected raters, and the reliability is calculated from a single measurement. Uncommonly used in clinical reliability studies.
- ICC(1,k) As above, but reliability is calculated by taking an average of the k raters' measurements.
- ICC(2,1) Each subject is measured by each rater, and raters are considered representative of a larger population of similar raters. Reliability calculated from a single measurement.
- ICC(2,k) As above, but reliability is calculated by taking an average of the k raters' measurements.
- ICC(3,1) Each subject is assessed by each rater, but the raters are the only raters of interest. Reliability calculated from a single measurement.
- ICC(3,k) As above, but reliability is calculated by taking an average of the k raters' measurements.

Summary ICC type of Shrout and Fleiss

13

- ICC(1,1) One-way random, single measure
- ICC(1,k) One-way random, average measure
- ICC(2,1) Two-way random, single measure
- ICC(2,k) Two-way random, average measure
- ICC(3,1) Two-way mixed, single measure
- ICC(3,k) Two-way mixed, average measure
- As a general rule, for the vast majority of applications, only 1 of S&F's ICCs [ICC(2,1)] is needed.

Single OR Average Agreement

14

- Two lots of ICC data are produced: one for the single measure, and one for the average measure. You decide which one to document based on the “form” of the ICC (whether you take a single measure or whether you average the measurements from multiple raters).
- Though it may be tempting to document the average measure (as it will be a better ICC), this is cheating unless you have decided a priori to use an average.
- In most cases, you will be using a single measure anyway.

Consistency OR Absolute Agreement (McGraw and Wong)

15

- For both 2-way random and mixed-effects models, there are 2 ICC definitions: “**absolute agreement**” and “**consistency**.”
- Selection of the ICC definition depends on whether we consider absolute agreement or consistency between raters to be more important.
- Absolute agreement concerns if different raters assign the same score to the same subject.
- Conversely, consistency definition concerns if raters’ scores to the same group of subjects are correlated in an additive manner.
- Consider an interrater reliability study of 2 raters as an example. In this case, consistency definition concerns the degree to which one rater’s score (y) can be equated to another rater’s score (x) plus a systematic error (c) (ie, $y = x + c$), whereas absolute agreement concerns about the extent to which y equals x .

Interpretation

16

- A low ICC could not only reflect the low degree of rater or measurement agreement but also relate to the lack of variability among the sampled subjects, the small number of subjects, and the small number of raters being tested.
- As a rule of thumb, researchers should try to obtain at least 30 heterogeneous samples and involve at least 3 raters whenever possible when conducting a reliability study.
- Under such conditions, we suggest that ICC values less than 0.5 are indicative of **poor reliability**, values between 0.5 and 0.75 indicate **moderate reliability**, values between 0.75 and 0.9 indicate **good reliability**, and values greater than 0.90 indicate **excellent reliability**.

How to Select the Correct ICC Form for Interrater Reliability Studies

17

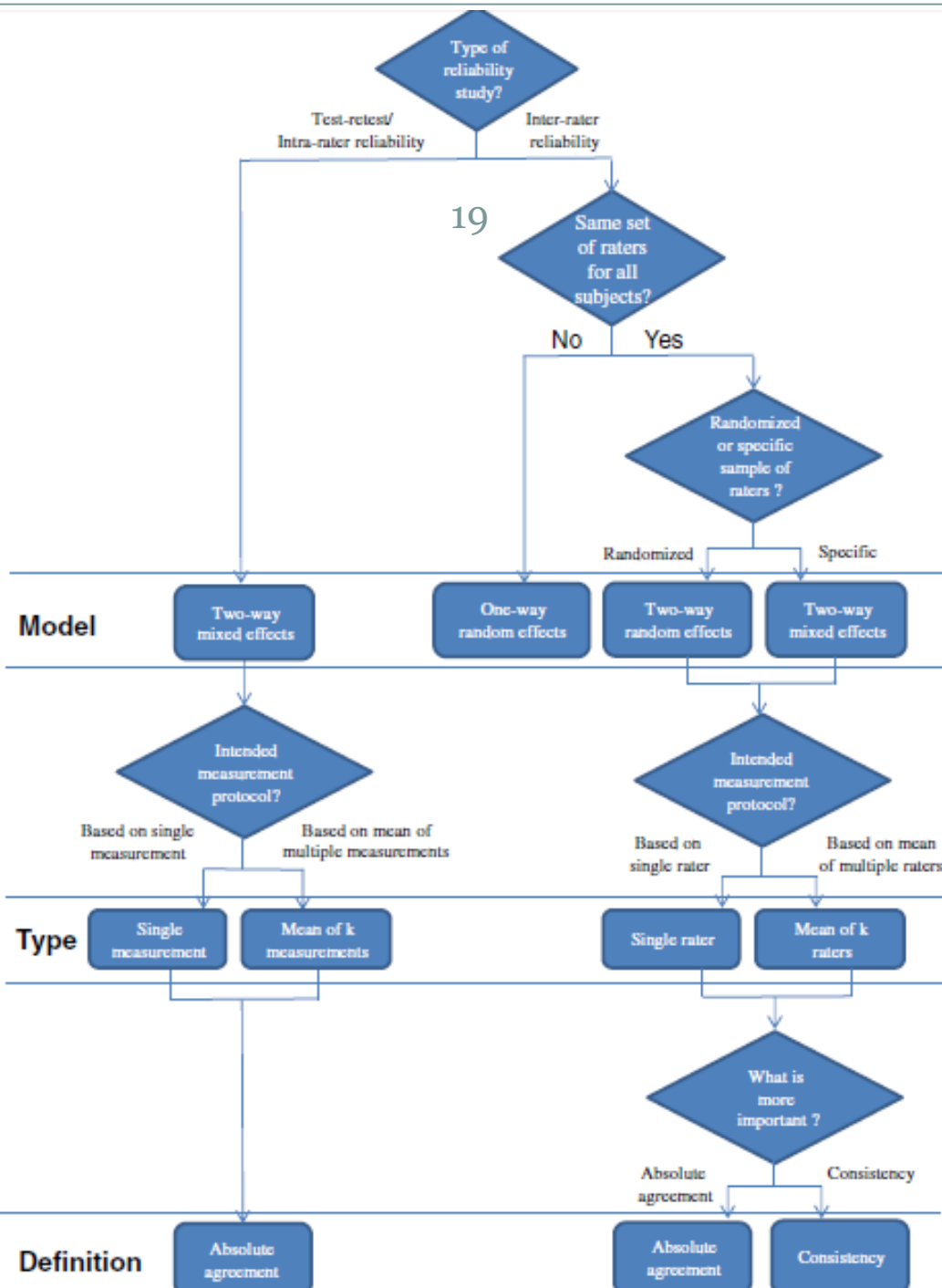
- Selection of the correct ICC form for interrater reliability study can be guided by 4 questions:
- (1) Do we have the same set of raters for all subjects?
- (2) Do we have a sample of raters randomly selected from a larger population or a specific sample of raters?
- (3) Are we interested in the reliability of single rater or the mean value of multiple raters?
- (4) Do we concern about consistency or agreement?
- The first 2 questions guide the “Model” selection, question 3 guides the “Type” selection,
- and the last question guides the “Definition” selection.

How to Select the Correct ICC Form for Test-Retest and Intrarater Reliability Studies

18

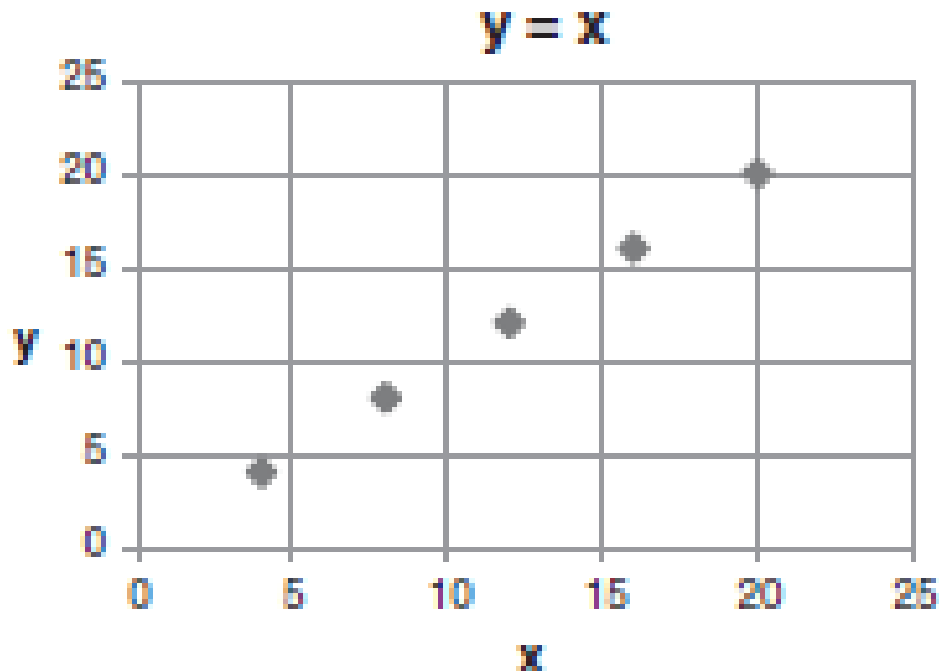
- Compared with inter rater reliability, the ICC selection process of the test-retest and intra rater reliability is more straightforward. The only question to ask is whether the actual application will be based on a single measurement or the mean of multiple measurements.
- As for the “Model” selection, Shrout and Fleiss suggest that 2-way mixed-effects model is appropriate for testing intrarater reliability with multiple scores from the same rater, as it is not reasonable to generalize one rater’s scores to a larger population of raters.
- Similarly, 2-way mixed-effects model should also be used in test-retest reliability study because repeated measurements cannot be regarded as randomized samples.
- In addition, absolute agreement definition should always be chosen for both test-retest and intrarater reliability studies because measurements would be meaningless if there is no agreement between repeated measurements.

19



ICC Characteristics

20



Single Measurement:

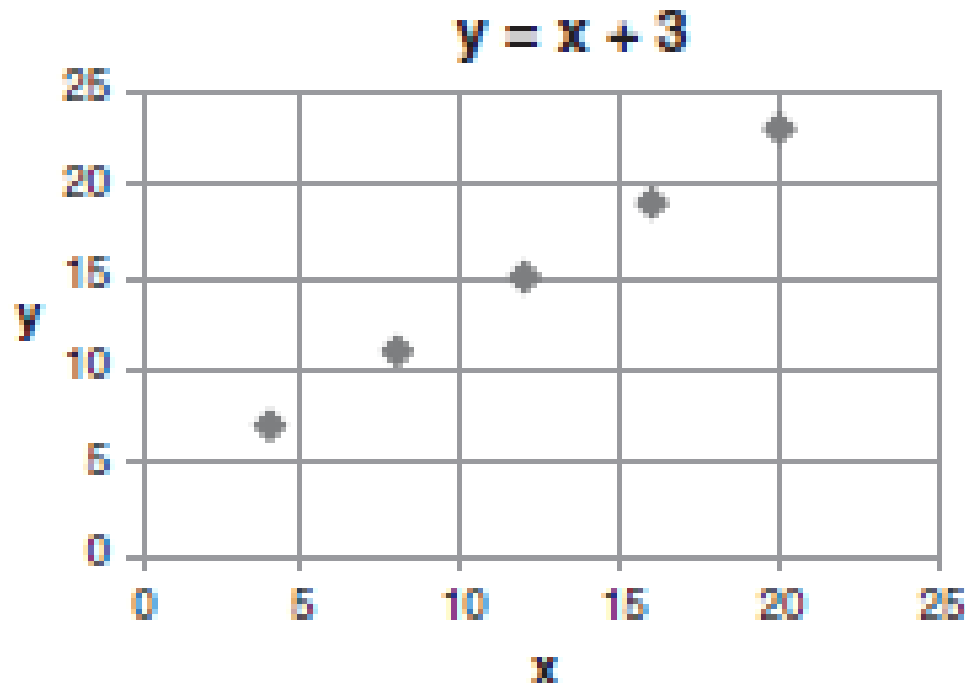
One-Way Random, absolute = 1.00
Two-Way Random, absolute = 1.00
Two-Way Random, consistency = 1.00
Two-Way Mixed, absolute = 1.00
Two-Way Mixed, consistency = 1.00

Mean Measurement:

One-Way Random, absolute = 1.00
Two-Way Random, absolute = 1.00
Two-Way Random, consistency = 1.00
Two-Way Mixed, absolute = 1.00
Two-Way Mixed, consistency = 1.00
Pearson $R^2 = 1.00$

ICC Characteristics

21



Single Measurement:

One-Way Random, absolute = 0.875
Two-Way Random, absolute = 0.882
Two-Way Random, consistency = 0.997
Two-Way Mixed, absolute = 0.882
Two-Way Mixed, consistency = 0.997

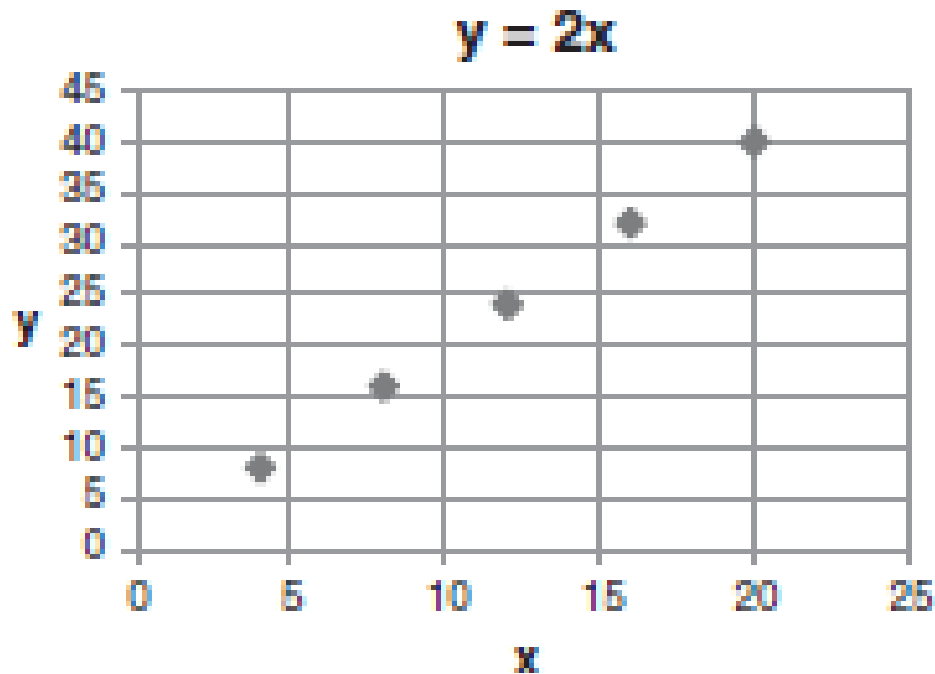
Mean Measurement:

One-Way Random, absolute = 0.933
Two-Way Random, absolute = 0.938
Two-Way Random, consistency = 0.999
Two-Way Mixed, absolute = 0.938
Two-Way Mixed, consistency = 0.999

Pearson $R^2 = 1.00$

ICC Characteristics

22



Single Measurement:

One-Way Random, absolute = 0.320
Two-Way Random, absolute = 0.461
Two-Way Random, consistency = 0.787
Two-Way Mixed, absolute = 0.461
Two-Way Mixed, consistency = 0.787

Mean Measurement:

One-Way Random, absolute = 0.484
Two-Way Random, absolute = 0.631
Two-Way Random, consistency = 0.881
Two-Way Mixed, absolute = 0.631
Two-Way Mixed, consistency = 0.881

Pearson $R^2 = 1.00$

ICC Characteristics

23

- (1) If the data sets are identical, all ICC estimates will equal to 1.
- (2) Generally speaking, ICC of the “mean of k raters” type is larger than the corresponding “single rater” type.
- (3) The “absolute agreement” definition generally gives a smaller ICC estimate than the “consistency” definition.
- (4) One-way random-effects model generally gives a smaller ICC estimate than the 2-way models.
- (5) For the same ICC definition (eg absolute agreement), ICC estimates of both the 2-way random- and mixed-effects models are the same because they use the same formula to calculate the ICC (Table 3). This brings up an important fact that the difference between 2-way random- and mixed-effects models is not on the calculation but on the experimental design of the reliability study and the interpretation of the results.

McGraw and Wong (1996) Convention ^a	Shrout and Fleiss (1979) Convention ^b	Formulas for Calculating ICC ^c
One-way random effects, absolute agreement, single rater/measurement	ICC (1,1)	$\frac{MS_R - MS_W}{MS_R + (k - 1)MS_W}$
Two-way random effects, consistency, single rater/measurement	24 —	$\frac{MS_R - MS_E}{MS_R + (k - 1)MS_E}$
Two-way random effects, absolute agreement, single rater/measurement	ICC (2,1)	$\frac{MS_R - MS_E}{MS_R + (k - 1)MS_E + \frac{k}{n}(MS_C - MS_E)}$
Two-way mixed effects, consistency, single rater/measurement	ICC (3,1)	$\frac{MS_R - MS_E}{MS_R + (k - 1)MS_E}$
Two-way mixed effects, absolute agreement, single rater/measurement	—	$\frac{MS_R - MS_E}{MS_R + (k - 1)MS_E + \frac{k}{n}(MS_C - MS_E)}$
One-way random effects, absolute agreement, multiple raters/measurements	ICC (1,k)	$\frac{MS_R - MS_W}{MS_R}$
Two-way random effects, consistency, multiple raters/measurements	—	$\frac{MS_R - MS_E}{MS_R}$
Two-way random effects, absolute agreement, multiple raters/measurements	ICC (2,k)	$\frac{MS_R - MS_E}{MS_R + \frac{MS_C - MS_E}{n}}$
Two-way mixed effects, consistency, multiple raters/measurements	ICC (3,k)	$\frac{MS_R - MS_E}{MS_R}$
Two-way mixed effects, absolute agreement, multiple raters/measurements	—	$\frac{MS_R - MS_E}{MS_R + \frac{MS_C - MS_E}{n}}$

More detail

25

- Two way random or mixed Single measure, consistency have the same formula (2,4)
- Two way random or mixed Single measure, absolute agreement have the same formula (3,5)
- Two way random or mixed Multiple measure, consistency have the same formula (7,9) .
- Two way random or mixed Average measure consistency= Cronbach's alpha
- Two way random or mixed Multiple measure, absolute agreement have the same formula (8,10)

How to Report ICC

26

- There is currently a lack of standard for reporting ICC in the clinical research community.
- Given that different forms of ICC involve distinct assumptions in their calculation and will lead to different interpretations, it is imperative for researchers to report detailed information about their ICC estimates.
- The best practice of reporting ICC should include the following items: software information, “Model,” “Type,” and “Definition” selections. In addition, both ICC estimates and their 95% confidence intervals should be reported.

Drawback of ICC

27

- The ICC is supported by the analysis of variance (ANOVA) .
- The main limitation of this method resides in its strong dependence on the variance of the assessed population. Higher ICC values may, thus, be obtained when applied to a more heterogeneous population as compared with a more homogeneous one despite similar levels of agreement .
- Consequently, the ICC values cannot be said to translate an absolute level of agreement, and the cutoff value of 0.75, proposed by Burdock et al. to signify good agreement

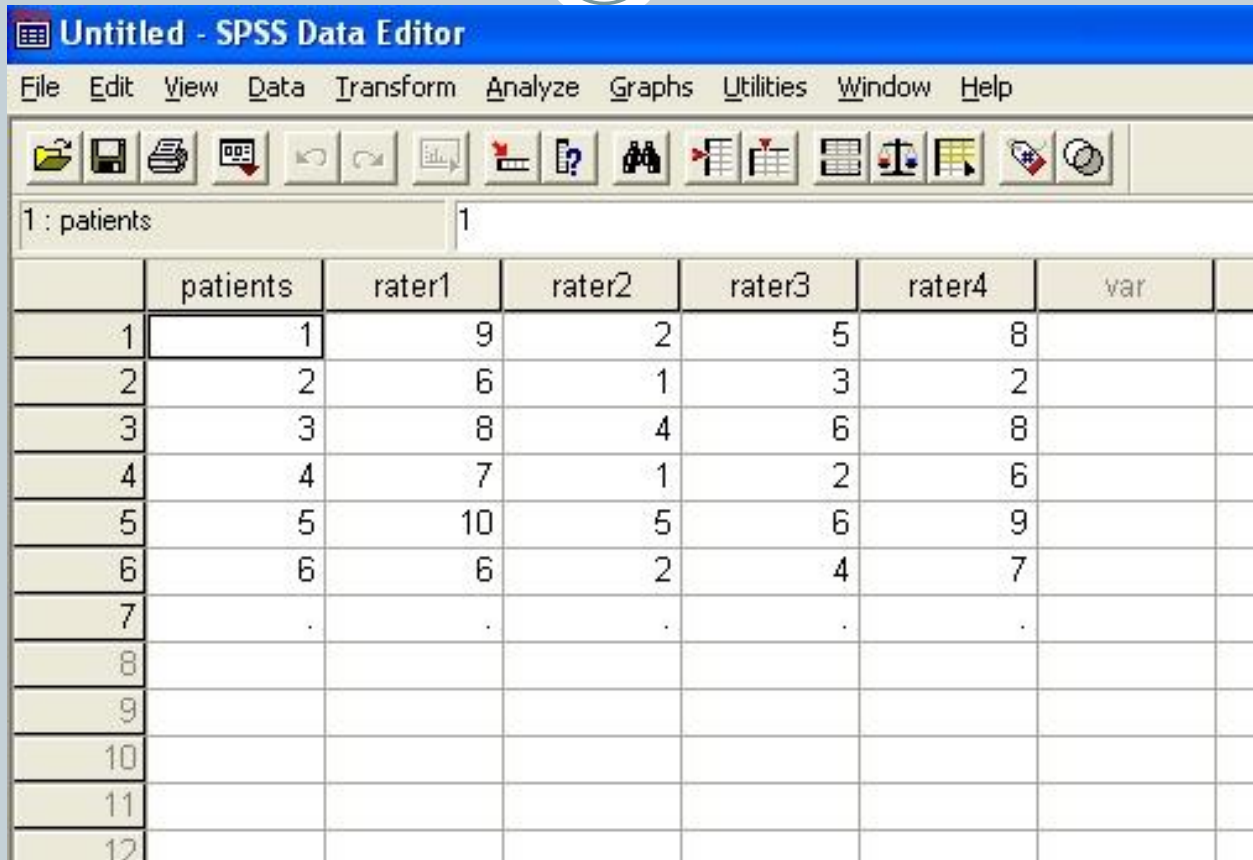
Example: Depression Ratings (Inter rater reliability)



Patients	Nurse1	Nurse2	Nurse3	Nurse4
1	9	2	5	8
2	6	1	3	2
3	8	4	6	8
4	7	1	2	6
5	10	5	6	9
6	6	2	4	7

Enter data into SPSS

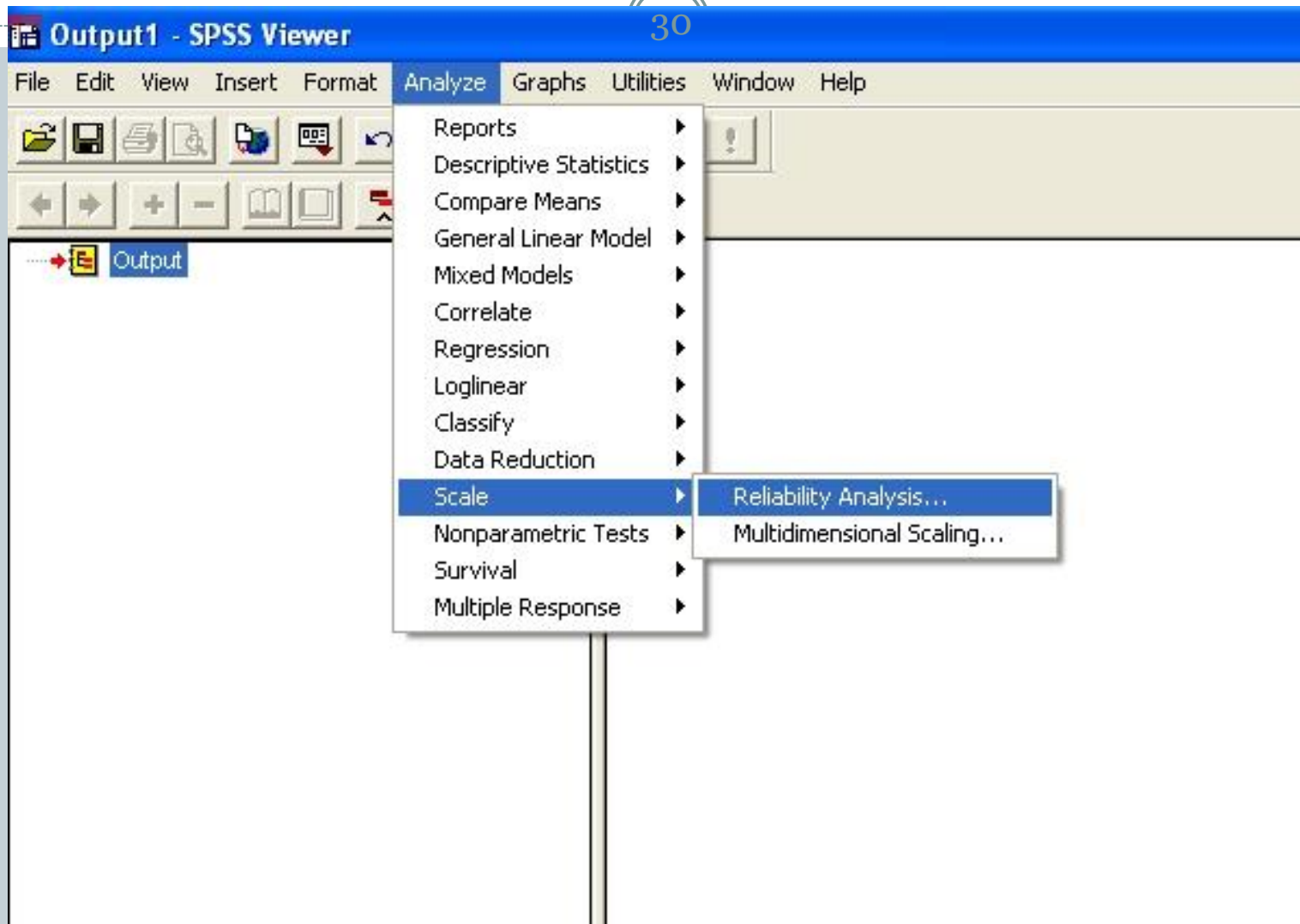
29



The screenshot shows the SPSS Data Editor window titled "Untitled - SPSS Data Editor". The menu bar includes File, Edit, View, Data, Transform, Analyze, Graphs, Utilities, Window, and Help. The toolbar contains various icons for file operations, editing, and analysis. The data grid shows 12 rows and 7 columns. The first column is labeled "patients" and contains values 1 through 6, followed by three empty rows and then rows 7 through 12. The second column is labeled "rater1" and contains values 9, 6, 8, 7, 10, 6, followed by three empty rows and then rows 7 through 12. The third column is labeled "rater2" and contains values 2, 1, 4, 1, 5, 2, followed by three empty rows and then rows 7 through 12. The fourth column is labeled "rater3" and contains values 5, 3, 6, 2, 6, 4, followed by three empty rows and then rows 7 through 12. The fifth column is labeled "rater4" and contains values 8, 2, 8, 6, 9, 7, followed by three empty rows and then rows 7 through 12. The sixth column is labeled "var" and contains empty cells for all rows. The seventh column is empty for all rows.

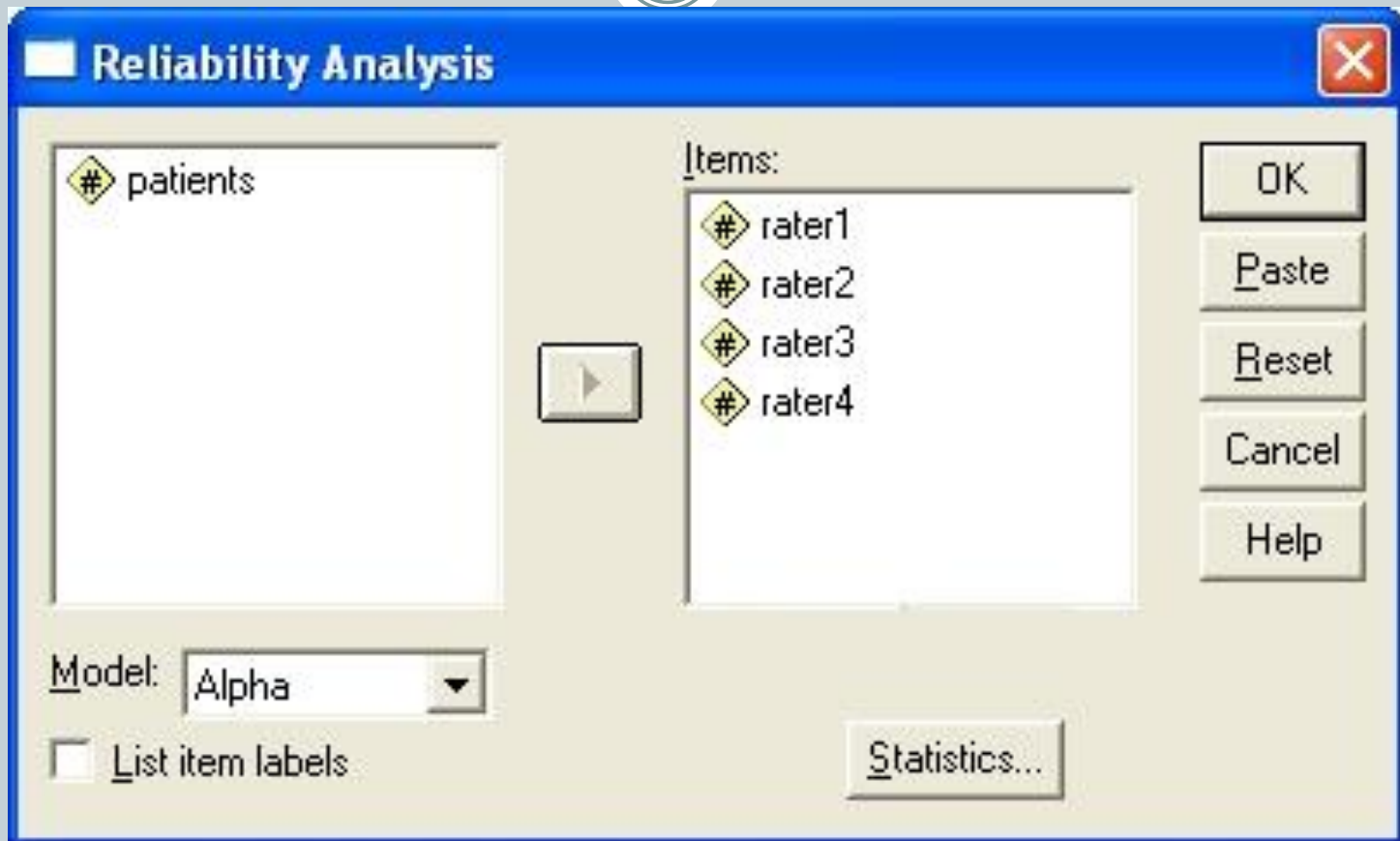
	patients	rater1	rater2	rater3	rater4	var
1	1	9	2	5	8	
2	2	6	1	3	2	
3	3	8	4	6	8	
4	4	7	1	2	6	
5	5	10	5	6	9	
6	6	6	2	4	7	
7	
8						
9						
10						
11						
12						

Find the Reliability Analysis



Select Raters

31



Choose Analysis

Reliability Analysis: Statistics

Descriptives for

☐ Item

☐ Scale

☐ Scale if item deleted

Inter-Item

☐ Correlations

☐ Covariances

Summaries

☐ Means

☐ Variances

☐ Covariances

☐ Correlations

ANOVA Table

☒ None

☐ F test

☐ Friedman chi-square

☐ Cochran chi-square

☐ Hotelling's T-square

☐ Tukey's test of additivity

☒ Intraclass correlation coefficient

Model: Two-Way Random

Type: Consistency

Confidence interval: 95 %

Test value: Consistency

Absolute Agreement

Continue Cancel Help

SPSS OUT PUT

Absolute Agreement

33

Intraclass Correlation Coefficient							
	Intraclass Correlation ^a	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.290 ^b	.019	.761	11.027	5	15	.000
Average Measures	.620	.071	.927	11.027	5	15	.000

Two-way random effects model where both people effects and measures effects are random.

a. Type A intraclass correlation coefficients using an absolute agreement definition.

b. The estimator is the same, whether the interaction effect is present or not.

Consistency

Intraclass Correlation Coefficient							
	Intraclass Correlation ^a	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.715 ^b	.342	.946	11.027	5	15	.000
Average Measures	.909	.676	.986	11.027	5	15	.000

Two-way random effects model where both people effects and measures effects are random.

- a. Type C intraclass correlation coefficients using a consistency definition-the between-measure variance is excluded from the denominator variance.
- b. The estimator is the same, whether the interaction effect is present or not.